Supplementary material for:

annotatr: Genomic regions in context

Raymond G. Cavalcante[1] and Maureen A. Sartor[1,2]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109; [2]Department of Biostatistics, University of Michigan

Supplementary Methods:
Where appropriate, genomic annotations include Entrez Gene IDs and Gene Symbols (genic and lncRNA annotations), and UCSC Transcript IDs (genic annotations) or ENSEMBL Transcript IDs (lncRNA annotations).

1. Construction of Genomic Annotations

CpG Features
We use the AnnotationHub R package (Morgan, 2016) to obtain the CpG island annotations when available (hg19, mm9, rn4, and rn5), and otherwise use the UCSC Golden Path (hg38, mm10, and rn6). Using functions in the GenomicRanges R package (Lawrence, 2013), we defined CpG shores as 2kb upstream and downstream of the CpG island boundaries and excluding CpG islands. CpG shelves are defined as the 2kb regions immediately upstream and downstream of the CpG shores opposite of the CpG island. Again, this excludes regions already annotated as CpG islands and CpG shores. See Figure S1A for a schematic of the CpG annotations, and Table S1 for the organisms and genome builds with CpG annotations.

Genic Features
We use the TxDb R packages for the specified genomes (e.g., TxDb.Hsapiens.UCSC.hg19.knownGene (Carlson, 2015) for human genome version hg19) and functions from the GenomicFeatures R package (Lawrence, 2013) to extract 1-5kb regions upstream of a TSS, promoters (<1Kb upstream of a TSS), 5'UTRs, exons, introns, and 3'UTRs. Intron/exon and exon/intron boundaries are defined as 200bp around the boundary. Intergenic annotations are taken to be the complement of the aforementioned annotations. We allow all genic annotations to overlap. See Figure S1B for a schematic of the genic annotations, and Table S1 for organisms and genome builds with genic annotations.

lncRNA Features
We use GENCODE long non-coding RNAs (lncRNA) from GENCODE at the transcript level (Harrow, 2006). For hg19 we use GENCODE v19, for hg38 we use GENCODE v23, and for mm10 we use GENCODE vM6. Relevant GENCODE biotypes (https://www.gencodegenes.org/gencode_biotypes.html) are included as part of the annotations.

Enhancer Features
We use enhancers defined via bi-directional CAGE transcription from the FANTOM5 consortium (Andersson, 2014) for human (hg19) and mouse (mm9). We provide enhancer annotations for hg38 and mm10 with the rtracklayer::liftover() function on the hg19 and mm9 enhancer annotations. Additional enhancer regions are defined within the chromatin state features (see below). Enhancers in hg38 and mm10 will be available in the April 2017 Bioconductor release, or users may download annotatr from the GitHub repository (https://github.com/rcavalcante/annotatr) to use this feature.

Chromatin State Features
We use the chromatin states given by chromHMM (Ernst & Kellis, 2012) in each of 9 human cell lines. The cell lines are: GM12878, H1-hESC, HepG2, HUVEC, HMEC, HSMM, K562, NHEK, NHLF. The genomic coordinates are with respect to hg19 only. In brief, numerous ChIP-seq experiments and a hidden Markov model were used to segment the genome into the following 15 functional chromatin

states: active promoter, weak promoter, inactive/poised promoter, strong enhancer (2 classes), weak enhancer (2 classes), insulator, transcriptional transition, transcriptional elongation, weak transcribed, polycomb repressed, heterochromatin, and repetitive/CNV (2 classes).

AnnotationHub Resources
Any GRanges class resource from the AnnotationHub R package can be converted to an annotatr annotation via the build_ah_annots() function. Some resources of special interest to users may be COSMIC mutations, GWAS catalog mutations, and ENCODE / Roadmap Epigenomics datasets. Among the ENCODE and Roadmap datasets are many transcription factor binding peaks and histone modification peaks. This feature will be available in the April 2017 Bioconductor release, or users may download annotatr from the GitHub repository (https://github.com/rcavalcante/annotatr) to use this feature.

2. Benchmarking with microbenchmark and lineprof
The microbenchmark R package (Mersmann, 2015) was used on three data sets to compare runtimes over 10 runs of annotatr v1.0.1, ChIPpeakAnno v3.8.1 (Zhu, 2010), and goldmine v1.0.0 (Bhasin & Ting, 2016). Results are reported in Figure S5 and Table S6.

Benchmarks were run on our lab server containing 40 cores and 128 GB of RAM. The three data sets, ranging in size from 31,000 to 2,500,000 lines, are:

1. A ~31,000 line ChIP-seq peak file from ENCODE for Pol2 in the Gm12878 cell line (ENCODE Consortium, 2012).
2. A ~290,000 line file of hydroxymethylation peaks resulting from macs2 (Zhang *et al.*, 2008) on GEO dataset GSE52945 (Figueroa *et al.*, 2010).
3. A ~2,500,000 line CpG bedGraph report from Bismark (Krueger & Andrews, 2011) on a whole genome bisulfite sequencing run (unpublished data).

Supplementary Table 1: A summary of annotations available for organisms and genome builds. Custom annotations may be used in conjunction with built-in annotations, or for organisms with no built-in annotations. Note, enhancers for hg38 and mm10 use the rtracklayer::liftOver() function on enhancers from hg19 and mm9, respectively.

Annotation Type	Organism	Genome Builds
Genic	Fly, Human, Mouse, Rat	dm3, dm6, hg19, hg38, mm9, mm10, rn4, rn5, rn6
CpG	Human, Mouse, Rat	hg19, hg38, mm9, mm10, rn4, rn5, rn6
lncRNA	Human, Mouse	hg19, hg38, mm10
Enhancers	Human, Mouse	hg19, hg38, mm9, mm10
Chromatin State	Human	hg19

Supplementary Table 2: Example of a BED6+ file used for input into annotatr. The BED6 format (https://genome.ucsc.edu/FAQ/FAQformat.html#format1) has 6 required columns in the following order: chr, start, end, name, score, and strand. Annotatr can interpret BED files with any number of columns after these 6 (the +), so long as they are named and their type is explicitly given (see ?annotatr::read_regions for details). The underlying rtracklayer::import() function can also read files that have the first 3, 4, or 5 columns. Additionally, bedGraph files are supported using the format='bedGraph' parameter. In this example file, the additional columns are used to provide the mean methylation levels of two groups of samples (mu0 and mu1) and the difference in percent methylation between them.

chr	start	end	DM_status	pval	strand	mu0	mu1	diff_meth
chr9	10849	10948	none	0.505	*	-10.73	79.98	90.71
chr9	10949	11048	none	0.223	*	8.72	86.70	77.98
chr9	28949	29048	none	0.553	*	0.07	0.12	0.05
chr9	72849	72948	hyper	0.012	*	44.88	72.46	27.58
chr9	72949	73048	none	0.175	*	17.76	28.44	10.68
chr9	73049	73148	hyper	0.029	*	3.80	4.14	0.34
chr9	73149	73248	none	0.280	*	1.62	2.21	0.59
chr9	73349	73448	none	0.190	*	-1.05	0.00	1.05

Supplementary Table 3: Example output of the annotate_regions() function as a GRanges object (A) and data.frame (B). (A) Output of GRanges object with extra columns containing extra data from the input regions (DM_status, pval, diff_meth, mu0, and mu1). In addition, a column giving complete details about the annotations is in the annot column, however the gene information is hidden in this output. Of note is that regions with multiple annotations are repeated (see rows 1-2 and 3-5). (B) Using data.frame() allows users to coerce this GRanges object into a flat table and expose the gene information (last five columns).

A

```
GRanges object with 79029 ranges and 6 metadata columns:
          seqnames                ranges strand |   DM_status       pval
             <Rle>             <IRanges>  <Rle> | <character>  <numeric>
      [1]     chr9         [10850, 10948]      * |        none  0.5045502
      [2]     chr9         [10850, 10948]      * |        none  0.5045502
      [3]     chr9         [10950, 11048]      * |        none  0.2227126
      [4]     chr9         [10950, 11048]      * |        none  0.2227126
      [5]     chr9         [10950, 11048]      * |        none  0.2227126
      ...      ...                    ...    ... .         ...         ...
  [79025]     chr9 [141098750, 141098848]      * |        none 0.21376652
  [79026]     chr9 [141103350, 141103448]      * |        hypo 0.04861602
  [79027]     chr9 [141103350, 141103448]      * |        hypo 0.04861602
  [79028]     chr9 [141108950, 141109048]      * |        none 0.26167927
  [79029]     chr9 [141108950, 141109048]      * |        none 0.26167927
           diff_meth       mu0       mu1                    annot
           <numeric> <numeric> <numeric>                <GRanges>
      [1] -10.732905  79.98192  90.71483           chr9:6987-10986:+
      [2] -10.732905  79.98192  90.71483              chr9:1-24849:*
      [3]   8.719527  86.70401  77.98449         chr9:10987-11986:+
      [4]   8.719527  86.70401  77.98449           chr9:6987-10986:+
      [5]   8.719527  86.70401  77.98449              chr9:1-24849:*
      ...        ...       ...       ...                       ...
  [79025] -12.726055  83.18445  95.91050 chr9:141074192-141107369:*
  [79026]  -4.035105  95.41078  99.44589 chr9:141101637-141105636:+
  [79027]  -4.035105  95.41078  99.44589 chr9:141074192-141107369:*
  [79028] -10.493345  84.89418  95.38753 chr9:141107681-141109733:+
  [79029] -10.493345  84.89418  95.38753 chr9:141107370-141109369:*
  -------
  seqinfo: 93 sequences from hg19 genome
```

B as.data.frame()

```
seqnames start   end width strand DM_status      pval    diff_meth       mu0
1     chr9 10850 10948    99      *      none 0.5045502 -10.73290471 79.981920
2     chr9 10850 10948    99      *      none 0.5045502 -10.73290471 79.981920
3     chr9 10950 11048    99      *      none 0.2227126   8.71952705 86.704015
4     chr9 10950 11048    99      *      none 0.2227126   8.71952705 86.704015
5     chr9 10950 11048    99      *      none 0.2227126   8.71952705 86.704015
6     chr9 28950 29048    99      *      none 0.5530958   0.07008468  0.124081
        mu1 annot.seqnames annot.start annot.end annot.width annot.strand
1 90.7148252           chr9        6987     10986        4000            +
2 90.7148252           chr9           1     24849       24849            *
3 77.9844878           chr9       10987     11986        1000            +
4 77.9844878           chr9        6987     10986        4000            +
5 77.9844878           chr9           1     24849       24849            *
6  0.0539963           chr9       26005     30004        4000            -
       annot.id annot.tx_id annot.gene_id annot.symbol           annot.type
1   1to5kb:34327  uc011llp.1     100287596      DDX11L5    hg19_genes_1to5kb
2     inter:8599        <NA>          <NA>         <NA>       hg19_cpg_inter
3 promoter:34327  uc011llp.1     100287596      DDX11L5 hg19_genes_promoters
4   1to5kb:34327  uc011llp.1     100287596      DDX11L5    hg19_genes_1to5kb
5     inter:8599        <NA>          <NA>         <NA>       hg19_cpg_inter
6   1to5kb:35839  uc011llq.1     100287171        WASH1    hg19_genes_1to5kb
```

Supplementary Table 4: Example of summarized information of a numerical column over the annotations. Shown is a subset of the result of the summarize_numerical() function by annotation types (annot.type) and the specific annotated regions (annot.id, an internal ID specific to annotatr) over the column containing change in percent methylation (diff_meth). The input regions are the results of tests for differential methylation as described in the text. Each row is an annotation and contains the average diff_meth (mean) and standard deviation (sd) over all the input regions intersecting the annotation (the total number of which is n). The annot.id column can be cross referenced with the annotated regions (Table S3) for information about the specific annot.id (such as Entrez ID or gene symbol) and the n intersecting input regions (such as the exact diff_meth values for each region).

annot.type	annot.id	n	mean	sd
hg19_genes_exonintronboundaries	exonintronboundary:301892	5	3.84	4.89
hg19_genes_introns	intron:282469	10	1.71	7.60
hg19_genes_introns	intron:287513	3	-2.55	3.07
hg19_genes_introns	intron:289069	4	0.93	7.61
hg19_genes_introns	intron:296414	2	13.89	4.67
hg19_genes_introns	intron:299213	3	-0.13	0.41
hg19_genes_promoters	promoter:35271	3	0.19	0.25
hg19_genes_promoters	promoter:37273	6	10.16	15.87

Supplementary Table 5: Example of summarized information of a categorical data column over the annotations. The summarize_categorical() function was used by type of annotation (annot.type) and differential methylation status (DM_status), a categorical data column defined as hyper, hypo, or none. The result indicates the number of annotated regions in each annotation type and with each of the DM_status types.

annot.type	DM_status	n
hg19_chromatin_K562-Insulator	hyper	66
hg19_chromatin_K562-Insulator	hypo	11
hg19_chromatin_K562-Insulator	none	394
hg19_cpg_inter	hyper	523
hg19_cpg_inter	hypo	596
hg19_cpg_inter	none	7052
hg19_cpg_islands	hyper	976
hg19_cpg_islands	hypo	50
hg19_cpg_islands	none	4621
hg19_cpg_shelves	hyper	63
hg19_cpg_shelves	hypo	70
hg19_cpg_shelves	none	1114
hg19_cpg_shores	hyper	477
hg19_cpg_shores	hypo	151
hg19_cpg_shores	none	2963
hg19_enhancers_fantom	hyper	100
hg19_enhancers_fantom	hypo	11
hg19_enhancers_fantom	none	497
hg19_genes_1to5kb	hyper	322
hg19_genes_1to5kb	hypo	91
hg19_genes_1to5kb	none	2334
hg19_genes_3UTRs	hyper	69
hg19_genes_3UTRs	hypo	31
hg19_genes_3UTRs	none	456
hg19_genes_5UTRs	hyper	191
hg19_genes_5UTRs	hypo	20
hg19_genes_5UTRs	none	1450

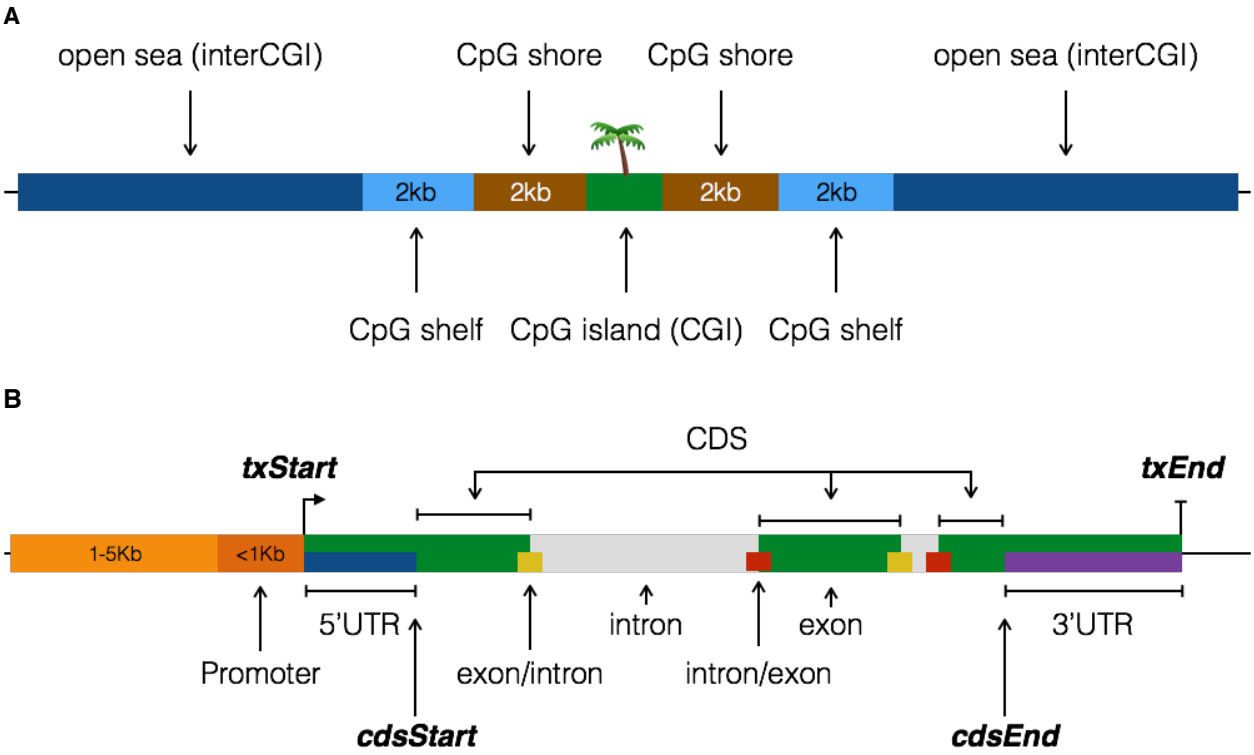
Supplementary Table 6: Benchmarking (in seconds, over 10 runs and 3 datasets) of ChIPpeakAnno and goldmine versus annotatr using the microbenchmark R package. In summary, the annotatr package tends to perform faster than competing packages.

File Size (lines)	Software	Runtime Min. (s)	Runtime Mean (s)	Runtime Max. (s)	X Mean / annotatr Mean
31k	ChIPpeakAnno	1.97	3.33	4.67	0.96x
	goldmine	9.31	11.17	12.79	3.2x
	annotatr	2.51	3.47	5.22	--
290k	ChIPpeakAnno	26.75	29.08	31.71	4.1x
	goldmine	46.55	52.92	58.16	7.51x
	annotatr	4.22	7.04	10.64	--
2.5m	ChIPpeakAnno	135.96	162.67	185.89	13.1x
	goldmine	318.56	341.11	375.75	27.5x
	annotatr	8.39	12.41	17.14	--

Supplementary Table 7: Feature comparison between comparable annotation tools.

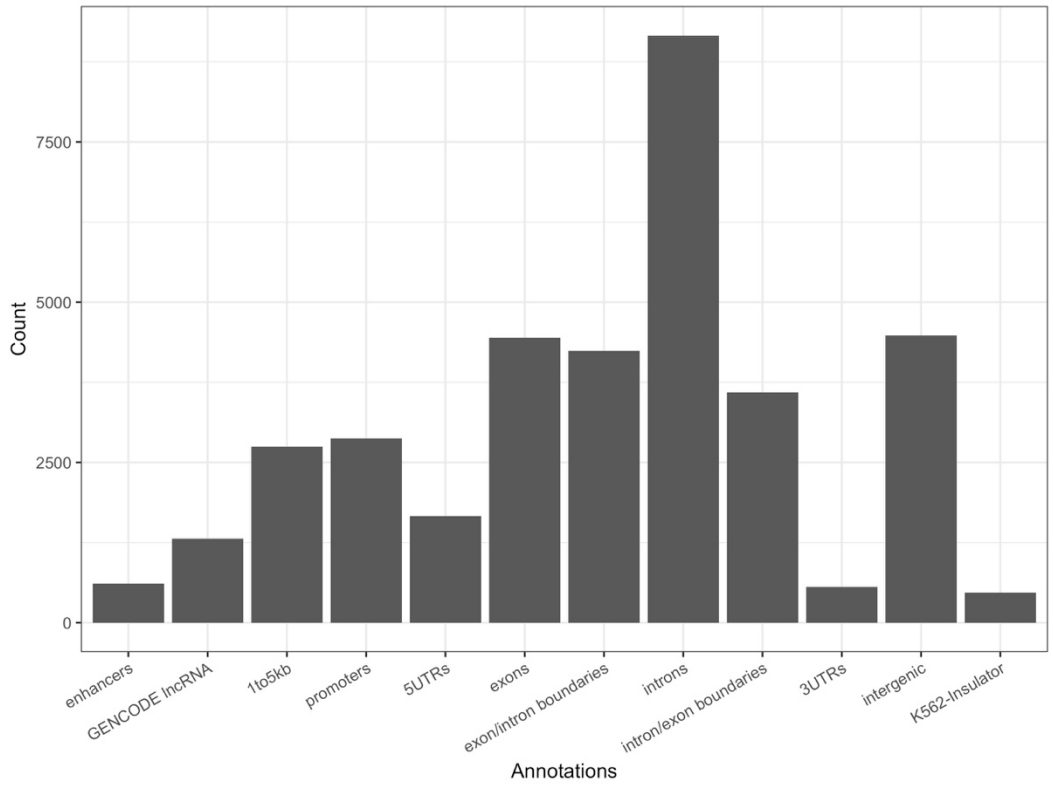
Feature	annotatr	goldmine	ChIPpeakAnno
Built-in Annotation Types			
CpG features	Yes	Yes	No
Genic features	Yes	Yes	Yes
A la carte selection of genic features	Yes	No	No
Enhancers	Yes	Yes	No
miRNA	No	Yes	Yes
lncRNAs	Yes	Yes	No
Chromatin States	Yes	Yes	No
Custom Annotations	Yes	Yes	Yes
Import Annotations from UCSC Tables	No	Yes	No
Annotation Reporting			
One-to-many annotation reporting	Yes	Yes	No
Prioritized annotation reporting	No	Yes	Yes
Summaries and Plots			
Summarization functions	Yes	Yes	Yes
Plot regions per annotation type	Yes	No	Yes
Plot regions per pair of annotation types	Yes	No	No
Plot region data over annotations	Yes	No	No
Plot region data over pairs of annotations	Yes	No	No

Supplementary Figure 1: Schematics of the CpG and genic annotation types used. (A) Schematic of the UCSC CpG annotations used in annotatr. The CpG islands are retrieved from either the AnnotationHub R package or the UCSC Golden Path, depending on availability for the genome build. CpG shores are defined as the 2kb extension upstream and downstream of the CpG island boundaries, less any CpG islands. The CpG shelves are a further 2kb extension upstream and downstream of the furthest upstream and downstream boundaries of the CpG shores, less any CpG island and shore annotations. The complement of the CpG islands, shores, and shelves make up the "open sea" or interCGI annotation. (B) A schematic of the genic annotations available in annotatr. Functions from the GenomicFeatures R package in conjunction with custom functions are used to extract regions 1-5Kb upstream of a TSS, promoters, 5'UTRs, exons, introns, and 3'UTRs. Additionally, exon/intron and intron/exon boundaries are determined by 200bp regions around such boundaries. Annotations may overlap one another from the same or from different transcripts. Genic annotations always have UCSC Transcript IDs and Entrez Gene IDs and gene symbols when applicable.

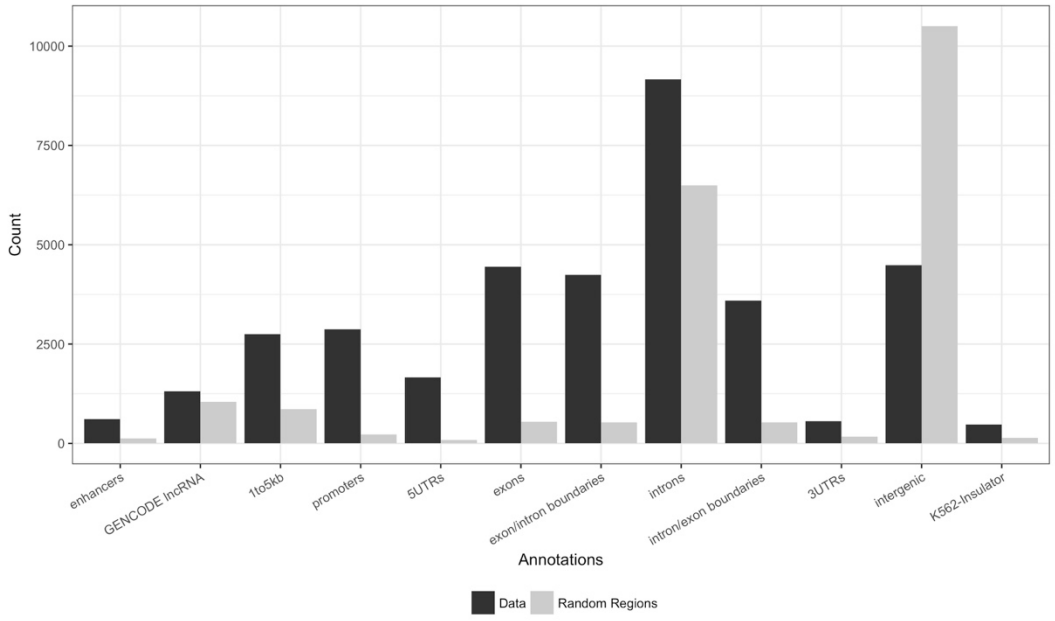


Supplementary Figure 2: Examples of annotatr barplots. The counts of regions per annotation type (A), and with annotations of random regions for comparison (B). In (B) we note that many annotations appear to be enriched (enhancers, promoters, exon/intron boundaries, and K562-insulators, and only intergenic regions are depleted. All plots are based on the ggplot2 package (Wickham, 2009).

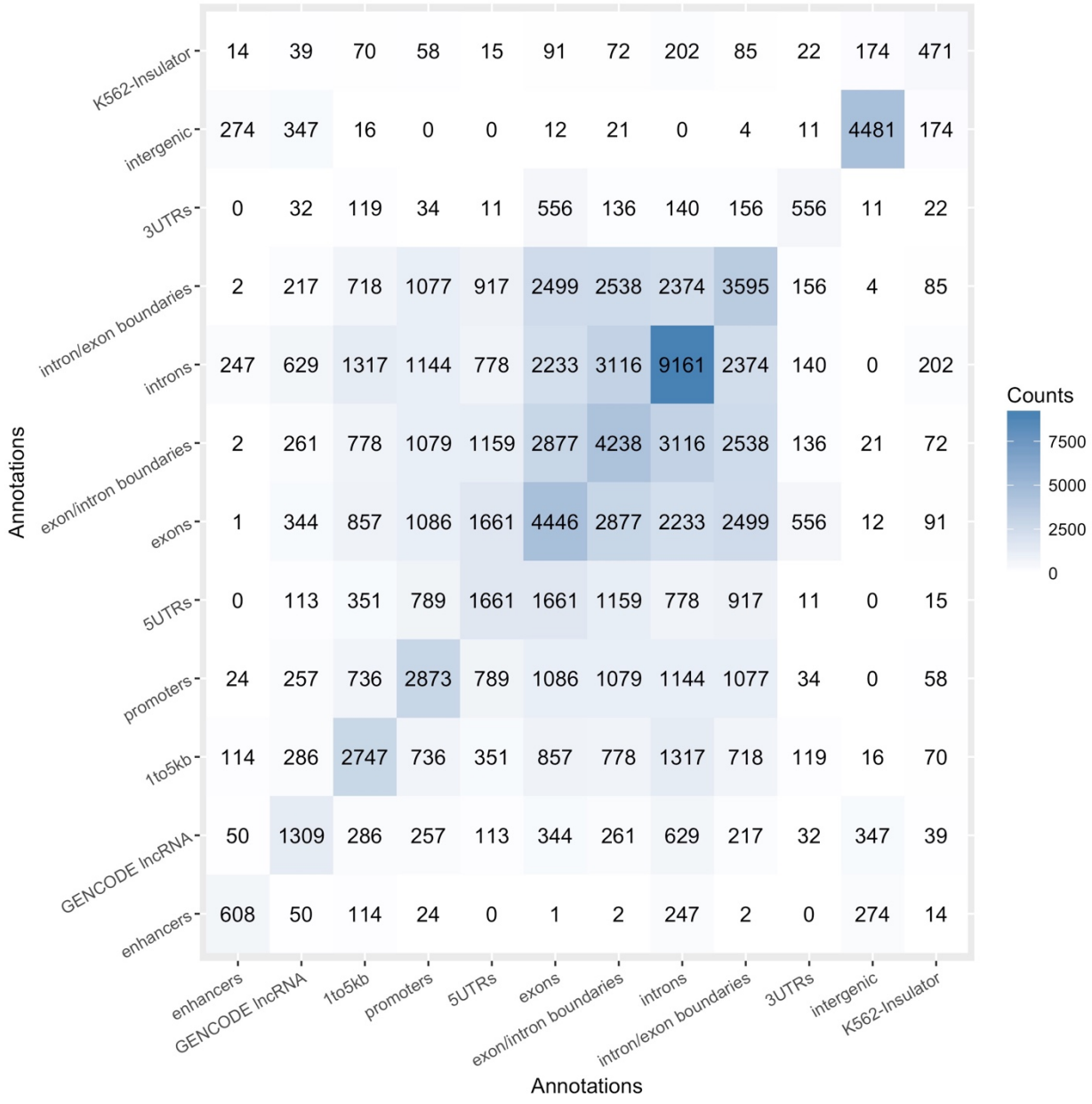
A



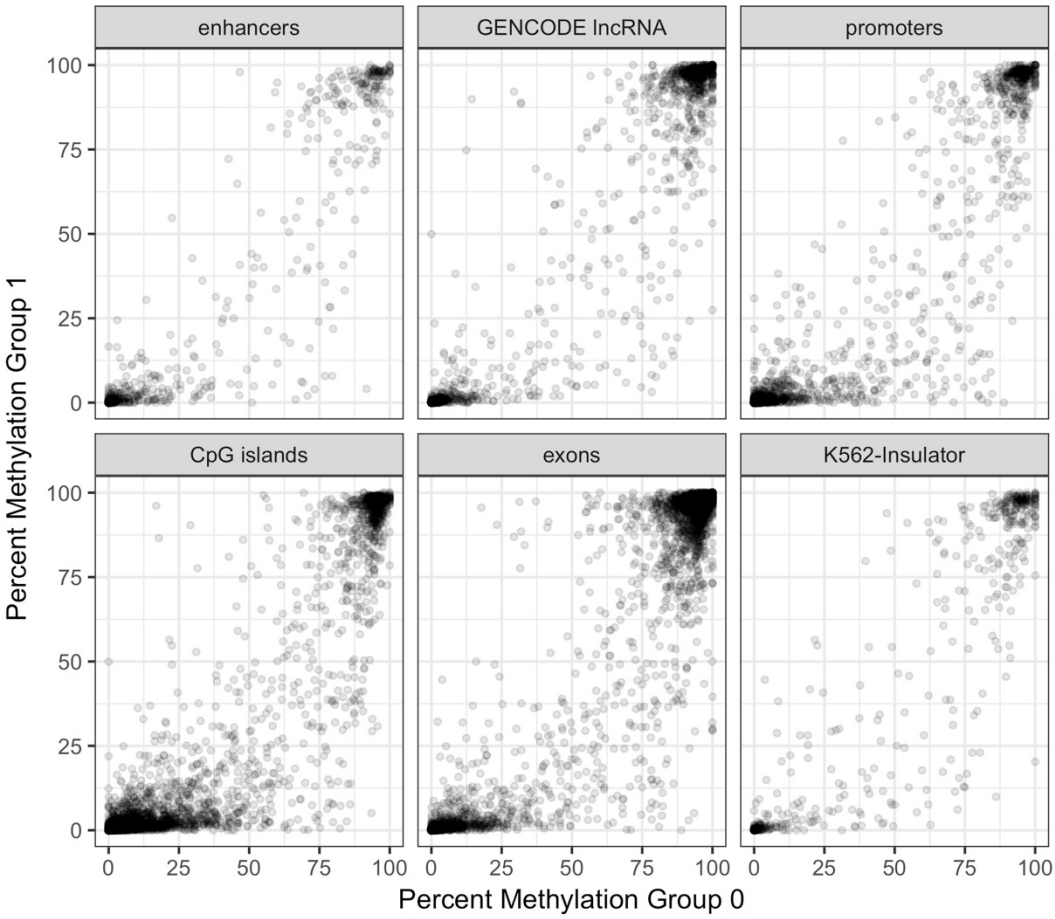
B



Supplementary Figure 3: The number of input genomic regions occurring in intersections of annotation pairs. This visualization is helpful for prioritizing types of regions to examine in more detail. For example, there are 247 regions that are in an enhancer and reside in an intron.

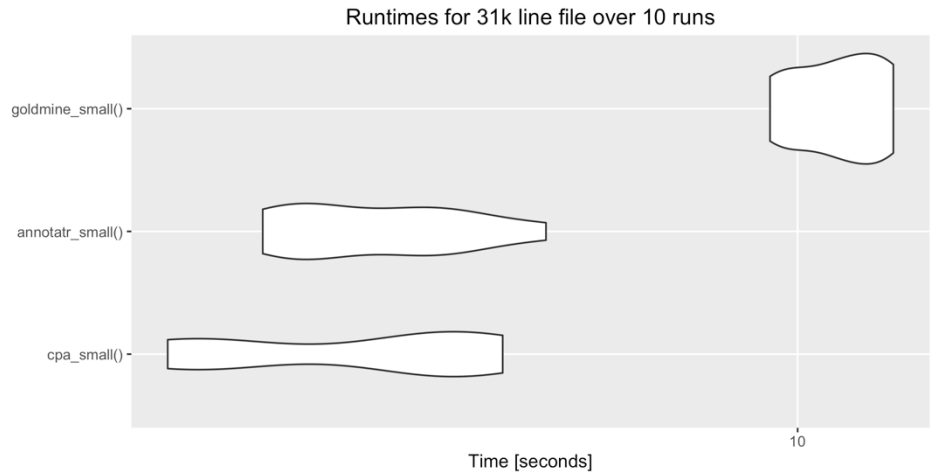


Supplementary Figure 4: Scatter plots of methylation rates comparing two sample groups across a subset of the annotation types. This visualization enables quick assessment of correlations in numerical data across different annotations types (or categorical variables).

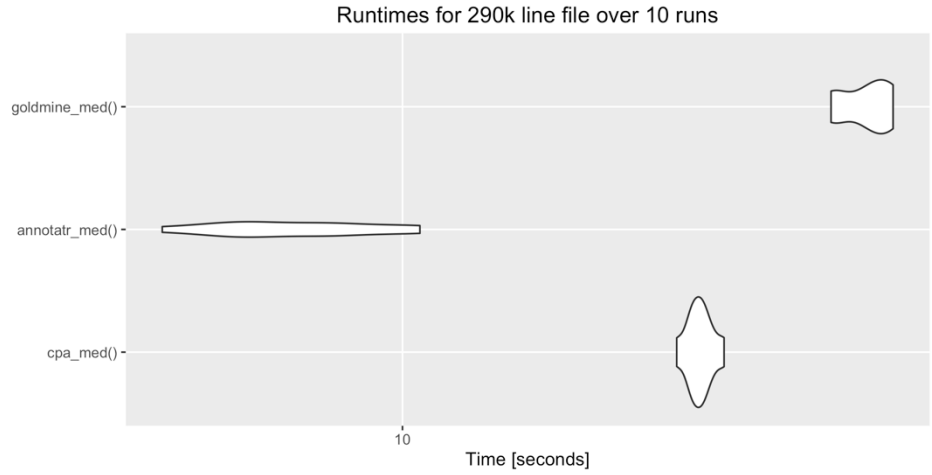


Supplementary Figure 5: Violin plots of benchmarking results comparing annotatr to ChIPpeakAnno and goldmine from file read to annotation for small (31k, A), medium (265k, B), and large (2.5m, C) files over 10 runs. Both annotatr and ChIPpeakAnno perform about the same for small files (A), but for larger files annotatr is clearly faster (B) and (C).

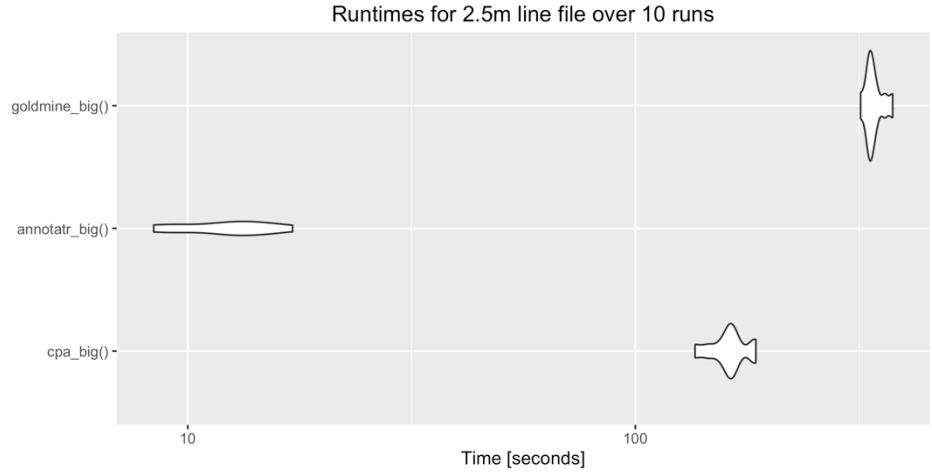
A



B



C



Supplementary References

Andersson, R., *et al.* (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*.

Bhasin, JM and Ting, AH. (2016). Goldmine integrates information placing genomic ranges into meaningful biological contexts. *Nucleic Acids Res*.

Carlson M and Maintainer BP (2015). *TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s)*. R package version 3.2.2.

ENCODE Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*, 57–74.

Ernst J and Kellis M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9, 215-216.

Figueroa, ME, *et al*. (2010). Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. *Cancer Cell*, *18*, 553–567.

Harrow J, *et al.* (2006). GENCODE: producing a reference annotation for ENCODE. *Genome biology*, 7 Suppl 1;S4.1-9

Krueger, F, & Andrews, SR. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, *27*(11), 1571–1572.

Lawrence, M., *et al.* (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, *9*(8), e1003118.

Mersmann, O. (2015). microbenchmark: Accurate Timing Functions. R package version 1.4-2.1. http://CRAN.R-project.org/package=microbenchmark

Morgan M., *et al.* (2016). *AnnotationHub: Client to access AnnotationHub resources*. R package version 2.6.2.

Wickham, H. (2009) ggplot2: elegant graphics for data analysis. Springer NY

Zhang Y, *et al*. (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biology, 9(9) pp. R137*

Zhu, LJ, *et al*. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, *11*, 237.